From the Digitized to the Digital Library [2001]

Thaller, Manfred

Historical Social Research
Historische Sozialforschung

Manfred Thaller:

From the Digitized to the Digital Library [2001]

doi: 10.12759/hsr.suppl.29.2017.307-319

Published in:

Historical Social Research Supplement 29 (2017)

Cite as:

Manfred Thaller. 2017. From the Digitized to the Digital Library [2001].
Historical Social Research Supplement 29: 307-319.
doi: 10.12759/hsr.suppl.29.2017.307-319.

Historical Social Research
Historische Sozialforschung

Other articles published in this Supplement:

From the Digitized to the Digital Library [2001]

Manfred Thaller[*]

Abstract: »*Von der digitalisierten zur digitalen Bibliothek*«. Based on a description of the major design decisions going into the Codices Electronici Ecclesiae Colonensis (CEEC) project, the role of such internet based digital collections in Humanities' research and beyond is addressed. From that discussion theses are submitted regarding: the communities which should be addressed by such collections and how to address them; the minimum size such collections should have; the quality used for the display of the digitized material; the possibilities for addressing objects persistently; the digital environment, into which the actual digital collection should be integrated; the role of such collections in academic teaching.

Keywords: Retrospective digitization, cultural heritage, manuscript processing, CEEC.

Background

The author holds a chair in Humanities Computer Science at the University of Cologne. For a number of years, he has been responsible for digitization projects, either as project director or as the person responsible for the technology being employed on the projects. The "Duderstadt project"[1] is one such project. It is one of the early large-scale manuscript servers, finished at the end of 1998, with approximately 80,000 high resolution documents representing the holdings of a city archive before the year 1600. The digital library of the Max-Planck-Institut für Europäische Rechtsgeschichte in Frankfurt[2] is another project on which the author has worked, with currently approximately 900,000 pages.

The author is currently project director of the project "Codices Electronici Ecclesiae Colonensis" (CEEC), which has just started and will ultimately consist of approximately 130,000 very high resolution color pages representing the complete holdings of the manuscript library of a medieval cathedral. It is being designed in close cooperation with the user community of such material. The project site[3], while not yet officially opened, currently holds about 5,000 pages and is growing by 100-

[*] Reprint of: Manfred Thaller. 2001. From the Digitized to the Digital Library. *D-Lib Magazine* February 2001, <http://www.dlib.org/dlib/february01/thaller/02thaller.html>.

[1] Formerly <http://www.archive.geschichte.mpg.de/duderstadt/dud-e.htm>; offline since 2013.

[2] <http://www.rg.mpg.de/bibliothek/digitale_bibliothek>.

[3] <http://www.ceec.uni-koeln.de>.

150 pages per day.[4] Parallel to the CEEC model project, a conceptual project, the "Codex Electronicus Colonensis" (CEC), is at work on the definition of an abstract model for the representation of medieval codices in digital form.

The following paper has grown out of the design considerations for the mentioned CEC project. The paper reflects a growing concern of the authors that some of the recent advances in digital (research) libraries are being diluted because it is not clear whether the advances really reach the audience for whom the projects would be most useful. Many, if not most, digitization projects have aimed at existing collections as individual servers. A digital library, however, should be more than a digitized one. It should be built according to principles that are not necessarily the same as those employed for paper collections, and it should be evaluated according to different measures which are not yet totally clear.

The paper takes the form of six theses on various aspects of the ongoing transition to digital libraries. These theses have been presented at a forum on the German "retrodigitization" program.[5] The program aims at the systematic conversion of library resources into digital form, concentrates for a number of reasons on material primarily of interest to the Humanities, and is funded by the German research council. As such this program is directly aimed at improving the overall infrastructure of academic research; other users of libraries are of interest, but are not central to the program.

2. Who Should be Addressed by Digital Libraries? How Shall we Measure whether we have Reached the Desired Audience?

Thesis: *The primary audience for a digital library is neither the leading specialist in the respective field, nor the freshman, but the advanced student or young researcher and the "almost specialist". The primary topic of digitization projects should not be the absolute top range of the "treasures" of a collection, but those materials that we always have wanted to promote if they were just marginally more important. Whether we effectively serve them to the appropriate community of serious users can only be measured according to criteria that have yet to be developed.*

Discussion:

[4] This site is still in the state of a beta test. Permanent accessibility is *not* guaranteed at the moment. Debugging messages of the underlying DBMS may appear in the dynamically created pages at short notice.

[5] <http://www.bsb.badw-muenchen.de/mdz/forum.htm>. In this English version of the paper given at the conference, we have cut down theses five and six to the bare minimum. They are either understandable only for the reader of German, as they relate directly to German language material, or are directly connected to current funding programs within Germany. Thesis four, which in our opinion is central to the future relationship between digital collections across the boundaries of national library systems and infoscapes, has been expanded.

Well-established academic staff have access to research assistants (RAs). Whether such RAs carry books to a copy machine or print from a screen is not an important concern. Well-established academic staff also usually have access to travel money. And academic travel, in most cases has intellectual and professional side effects, including personal contacts with academic colleagues at institutions visited, that go considerably beyond the value of library resources read during such travel. It is an illusion, therefore, to assume that even the most advanced and elaborate digital systems will entice researchers to stay at home when they can afford to visit collections where they can be sure to meet colleagues whom they would not otherwise meet.

Similarly, the well-established specialists within a field will usually already have large collections of copies of all but the most marginal types of resources. Digitization of resources that they already know cannot serve them. Every researcher, however, is usually aware of a large number of library resources that are relatively peripheral to his or her primary concern, but which *would* be important if they could be accessed easily. Therefore, making *such* material accessible digitally will increase the opportunities of a researcher.

As a corollary, resources that have been edited in print in the 18th century, lithographed at the beginning and photographed at the end of the 19th century, rephotographed in color in the middle of the 20th century, and reproduced as high end facsimile towards the end of the century are *not* the appropriate subjects of digitization. Materials that always had to be left out of past reformatting projects because it was too expensive to include them are appropriate subjects for digitization.

When we abandon the most obvious strategy, however, of digitizing well-known and frequently reproduced items, we need to find serious and reliable criteria to evaluate the digitization project's success. Nobody has to justify the digitization of the *Magna Carta*, even if nothing is gained by it (since its text is already ubiquitous). But who actually profits, and by how much, if we provide a digital version of the less well-known *cartae* of the same period?

As a first step in developing new evaluation criteria, we should drop the most obvious mark of success. There are few things in the world that are as totally irrelevant and meaningless for the success of a digitization project, as the raw hit rate at its server. Hit rates are, to be precise, as meaningful for a library as the number of tourists gaping at the glorious murals in the entrance hall: a good way to make friends, but no real guarantee that the local chamber of commerce will not propose to close down the costly remainder of the building.

In one of the projects with which the author has been involved, we have quite carefully analyzed the logs of the server, which boasted some five- or six- digit hit rates. When we looked more closely, we tried to decide what exactly a criterion for "real use" could be. We assumed that the only people who could be considered to be "real readers" were those who accessed a minimum of three successive pages in a book. Furthermore, there had to be a sufficient temporal interval between each page access so that it was plausible that each page was actually read and understood. Using these criteria reduced the spectacular numbers quite drastically; this special collection may have gained only one serious user per day, however, who typically reads for a few hours. This should be augmented with at least two more

users each day who, according to the logs, have consulted the transcribed tables of content systematically, but have not qualified as "serious readers" under the rather rigid criterion applied above. This may come as a shock, if we compare it with the "countless hits" which are usually quoted in such cases.

But these are real *users*, some of whom would have had to travel from Japan, Turkey or the US to Germany for a number of weeks to consult the same literature. And if you compare the travel costs incurred by the (fictitious) international research community to fund the travel, digitization projects suddenly become less expensive than they may look otherwise. One of the problems with this argument may be that in many cases those profiting the most from digitization of research material are, in the short run, members of research communities other than the community that funded the project. Fortunately, German research funding traditionally has been reasonably un-parochial in its perspectives.

Recommendation: *The ongoing digitization projects should systematically develop clear and open criteria for the actual usage of the resources created. These criteria should be conservatively defined. Evaluations of the cost effectiveness of a digitization project should compare the cost of digitization with the costs incurred if access to the material was by other means.*

2. The Appropriate Size of Digital Libraries and their Access Tools

Thesis: *Digital collections need a critical, minimal size to make their access worthwhile. In the end, users want to access information, not metadata or gimmicks.*

Discussion:

It is a well-known truism that the Internet as a whole creates an information glut. Therefore, a corollary goes, one of the primary challenges for the evolving worldwide infoscape is to define means to prevent the user from being overwhelmed by the flood. This in turn leads to the further claim that the primary challenge for digital libraries is the development of metadata standards.

For libraries this is undoubtedly true. To interconnect the OPACs of a national or trans-national library system, agreement about descriptive standards is obviously central. From this baseline, however, conclusions have been drawn for digital collections that may easily turn out to be very counterproductive.

Some background assumptions, before we go on:

The creation of digital collections does not *have* to be particularly expensive anymore. One of the more spectacular technical developments in recent years has been the drop in the pricing of digital cameras, where the resolution achievable by a $1,000 camera has been climbing sharply. At the other end, cameras like the 4096 x 4096 pixel camera offered by Kodak, with an observed workflow of ca. 5 / 10 seconds per exposure, are today still in the six-digit price range. With an emerging mass market of digital hobby photographers, it seems to be a safe bet that high speed digital cameras at a professional resolution will become achievable for rou-

tine projects in less than ten, presumably within the next five years.[6] This means that with 2,000 exposures per campaign day – roughly the speed possible with analog cameras today, the handling of the object being a serious barrier for quite some time – 1,000,000 page digitization projects will be possible with a limited budget and over a two-year or 500-day time frame.

For reasons that are beyond the scope of this paper and which have to do with the intricacies of the impact of the IT revolution on management hierarchies, the organization of document servers is today still believed to be a major technical challenge, an assumption understandably supported by the software industry profiting from it. The author would, however, like to emphasize that the reference systems he is involved with have all been created at rather low cost and within very short periods. Indeed, part of the training plans at the University of Cologne aims to bring the requirements for the creation of digital libraries to a level where the implementation of a digital library of the technical scope of the reference systems quoted initially, can be given within the next 12 months as a seminar assignment.

We propose, therefore, to base consideration of the required access tools for digital collections on the assumption that one million page collections can be produced reliably and cheaply within the near future.

One million pages seem like a lot. In the case of printed books, however, that number of pages represents a collection of something like 2,000-3,000 volumes. This is substantial enough that a user will profit from such a collection, and therefore be willing to learn how to use it. On the other hand, 2,000-3,000 volumes are usually not random collections of information with arbitrary levels of authority, quality and subject matter, as would be the case with an equivalent 1,000,000 web pages. If a researcher is interested in the development of religious doctrine in the early eighteenth century, he or she will probably be very willing to increase his or her understanding of the matter by browsing through a reasonably pre-ordered library of volumes with an overall relationship to the subject. If somebody, as it happens, should *not* be interested in early eighteenth century religiosity, even the most elaborate access system based on a highly sophisticated markup system will probably not seduce her or him into the depths of the collection.

Collections of this size would be eminently useful for research. And a collection of such size does *not* need to be made accessible beyond the levels of metadata that traditional library catalogues provide. A million-page collection with catalogue metadata is, to be precise, considerably more useful than a collection of ten volumes with complete transcriptions encoded according to an elaborate markup scheme developed for the occasion.

All of the above should, of course, be seen under a *mutatis mutandis* reservation. In the case of classical antiquity, one million pages of text probably come pretty close to the complete corpus of surviving texts. In a large number of cases, important ones, maximum accessibility of a relatively small number of items will be

[6] "Professional resolution" in this context means the resolution required to produce a digital object that can replace the functionality of the original on a screen today, and has a sufficient quality reserve to remain useful for the foreseeable future, i.e., 20-30 years.

extremely worthwhile. And, of course, there exist special collections where a smaller number of digital items constitutes an important tool. But one has to emphasize that there exist even more cases where one of the primary promises offered by digitization is the ability to have access to large bodies of material beyond what is accessible today.

For collections in which increased access to the content matters the most, the data itself, and not the metadata, are important. The further creation of small pilot projects with shinning interfaces that lure the user in, but which later on frustrate him or her because the project does not contain practically useful amounts of data, is increasingly detrimental to the acceptance of digital libraries as serious tools.

Ultimately, it is the content of a library that counts, not the architecture of the building housing it.

Recommendation: *A model for the creation of digital collections should be developed that allows for the creation of digital libraries in the one million primary digitization object (roughly: pages) range at minimal cost. Costs can most easily be minimized by cutting down on the effort invested in the creation of access information for the individual item.*

3. The Quality of Digital Objects

Thesis: *If digital library resources are to be integrated into the daily work of the research community, they must appear on the screen of the researcher in a quality that is useful in actual work.*

Discussion:

Digital objects can be characterized by the degree to which they allow functional replacement of the original. Four levels of digital objects can be differentiated. The names of the four levels have been derived from discussions around the creation of manuscript servers.

A digital object is called *illustrative* if its quality is sufficient to allow a user to make an informed decision about whether access to the original is worthwhile. This level of digital quality is usually employed by museum systems, since the impression of an original piece of art still goes beyond any impression that can be created on any screen available to the humanities research community.

A digital object is called *readable* if its quality allows the user to access all the information that the creator of the original object wanted to convey to the user. Digitized pages of a printed book, for example, have to be clearly readable on the screen and not strain the eyes. It is not necessary, however, to be able to decipher the notes that a few generations of college students in final examination frenzy have left in the margins.

A digital object is called *paleographic* in our terminology if the quality allows the user to access all the information that can be derived from the original with the unaided eye. In medieval codices it is important to be able to read the text. It is also important, however, to be able to see if in the lettering there is a recognizable change in the way the pen was held, thus indicating a change of authorship.

Finally, a digital object is *enhanceable* if the digital version provides access to information that cannot be extracted from the original with the unaided eye. Image enhancement may, for example, make erasures legible again.

Any quality much below "readable" is pointless in library applications. Thumbnails, for example, are eminently useful in museum systems, but are rarely useful by themselves in library systems.

If end user requirements, rather than absolute numbers, determine the appropriate quality of images in a digital system, it follows that a digital library has to specify – and discuss – a specific platform it expects its users to have, a specific purpose they are expected to follow in looking at the material, and the actual resolutions derived from these assumptions.

The following model for a manuscript library is offered for discussion:

Assumptions:

a) Professional manuscript work cannot be done on screens with a resolution of less than 1024 x 768. No specific support is given for analytic work on screens below that resolution.
b) 1024 x 768, however, defines the lower limit. 1200 x 1024 is considerably superior as soon as we go beyond the plain reading of manuscripts.

From these assumptions the following resolutions, with corresponding examples, have been derived:

The lowest resolution, which is displayed while browsing or searching within the framework of the main interface of the digital manuscript library, is defined as *Visual summary*.[7] It is high enough that a meaningful decision can be made regarding whether one of the higher resolutions should be loaded. It also provides, in exceptional situations, some support for 800 x 600 screens.

For standard work, however, two higher resolutions are provided: *Working copies*[8] are at a resolution that presents the horizontal dimension without scrolling on 1024 x 768 screens and preserves most of the optical properties of the original. *Optimized working copies*[9] are at a still higher resolution, and present their horizontal dimensions without scrolling on 1200 x 1024 screens. The area of the page that actually contains the writing will in most cases also fit into the horizontal dimension of 1024 x 768 screens. These pages are optimized in a mechanical way, i.e., some contrast enhancement and similar operations have been applied that optimize the readability as far as possible, without analyzing the characteristics of the pages individually. The price to be paid for this optimization is a distortion of, in particular, the colors.

For rare cases of detailed professional work, specifically in the area of paleography, a pretty *high resolution*[10] image, close to 4491 x 3480 in size, is presented. We are proud to bring that resolution to the world of complete digital collections, which

[7] <http://www.ceec.uni-koeln.de/ceec-cgi/kleioc/0010/exec/pagesma/%22kn28-0083ii_164.jpg%22>.
[8] <http://www.ceec.uni-koeln.de/ceec-cgi/kleioc/0010/exec/pagemed/"kn28-0083ii_164.jpg>.
[9] <http://www.ceec.uni-koeln.de/ceec-cgi/kleioc/0010/exec/pagepro/"kn28-0083ii_164.jpg>.
[10] <http://www.ceec.uni-koeln.de/ceec-cgi/kleioc/0010/exec/pagebig/%22kn28-0083ii_164.jpg%22>.

has so far been available only for CD-ROM-based facsimiles of individual manuscripts, as in the case of the Beowulf project.

Recommendation: *Digital libraries that do not offer the resolutions needed for professional work are useless. An explicit definition of the qualities offered, based on discussions with the potential customers, should therefore be included in the specifications of all digital resources*.

4. The Granularity / Modularity of Digital Repositories

Thesis: *While digital libraries are self-contained bodies of information, they are not the basic unit that most users want to access. Users are, as a rule, more interested in the individual objects in the library and need a straightforward way to access them*.

Discussion:

Digital libraries, particularly large ones, are still seen today as unique and significant projects. As a result, they are frequently constructed as self-contained systems, where the separation between the interface of the library and its contents is not as clear-cut as one would wish. This means that many digital libraries expect that a user will enter through the interface of the library. This is an example of when the implementation of a traditional metaphor is counterproductive.

In our reference project, we are experimentally creating a functionally complete linkage interface that allows one to access the content of the library completely independent of its own user interface. While this specification is not yet fully stabilized and public, it is partially available, and the following ways of addressing are guaranteed to be as persistent as the floating discussion of persistent basic identifiers allows. A researcher who intends to refer to the content of the Cologne manuscript library will have a mechanism that allows him or her to address reliably and persistently the following:

1) A digital object that represents a conventional unit of reference within a given discipline. In our case it is a medieval codex.
2) A digital object that represents the same object at a finer level of granularity, reflecting the usage of a given discipline. In our case individual pages are at the finer level of granularity.

Note: We refer intentionally to “units of references“ and “granular objects“ instead of “codices“ and “pages,“ not to introduce an additional level of complexity, but to prepare for the generalization of such addressing schemes to other cultural heritage material. Ultimately codices can be seen as particularly simple cases, where only one level of subdivision exists and the granular objects are ordered linearly, as opposed to, e.g., museum objects, where a number of hierarchical levels for digitization of details exist, and intuitive schemes for the naming of granular entities are considerably more complex. The *basic* problems remain the same, however.

The two types of reference above are necessary for two reasons:

3) From the end user's point of view, it is important to be able to include a reference to a digitally stored manuscript directly in a text. This will become much more important in the future when the results of research are themselves pre-

sented on digital media. In such cases it would be almost absurd to have an end user directed from a footnote to the search engine of a digital library, instead of the digital object itself, the address of which was obviously accessible to the author at the time of writing.

4) From the conserving institution's point of view, a clear tendency towards virtual libraries / archives / museums seems to exist. The most obvious way to construct such virtual collections is to envisage them as access platforms that hide from users the fact that the individual objects accessed are stored under different administrative and technical conditions. This is achieved most easily if an access machine can access individual digital objects in different holdings directly, that is, without a negotiation process with the access tools of the specific institution holding the object.

It would be highly impractical to rely on a central body, operating worldwide, to create a new set of identifiers for all existing objects of cultural heritage. All existing collections of manuscripts, archives, museums, etc., would have to agree upon a common system of shelfmarking for their objects. This is not only impractical, but also directly damaging, as the reference systems within collections of cultural heritage material that have grown historically usually represent by themselves a specific intellectual view of that material.

We envisage, therefore, a solution that divides the general problem into three sub problems:

1) A persistent addressing scheme for *collections*, which by necessity must be organized nationally, with national (or regional) solutions being coordinated by appropriate international bodies.
2) A persistent addressing scheme for digital objects within individual collections that is under the control of the individual institution, but which guarantees a common functionality and interoperability of the different collections.
3) A mapping scheme that allows referencing a granule of a digital object by a specific numbering scheme, which is then translated into the actual names of individual digital components, like page images. Such a mapping scheme is administered by the individual collection and should even exist if the names of the digital objects – file names – also reflect the traditional references directly. The order of access to granules of digital objects – "next page", for example – is a matter of interpretation. To allow operations like "virtual rebinding" of a digital codex, we strongly propose to differentiate clearly between this level and the preceding one.

An implementation of an addressing scheme for digital objects based on the preceding analysis would look as follows:

<collection-reference> <object-reference> <granule-reference>

where <granule-reference> is *either* a <direct-granule-reference> *or* a <mapped-granule-reference>.

CEC has a processing model for <object-reference> and <granule-reference>. For the discussion / definition of the concept of a <collection-reference> we seek the support of appropriate library institutions. CEEC has a working implementation for <object-reference> and <direct-granule-reference>, and a working implementation for the concept of a <mapped-granule-reference> is expected soon.

A complete Cologne codex can currently be reached via a WWW address like: http://www.ceec.uni-koeln.de/ceec-cgi/kleioc/0010/exec/katk/%22kn28-0083ii%22.

Ignoring the "%22" (the CGI wrap-up for the quotation marks), this means our previous definitions are realized as follows:
<collection-reference> = http://www.ceec.uni-koeln.de
<object-reference> = ceec-cgi/kleioc/0010/exec/katk/%22kn28-0083ii%22
To access an individual page of a Cologne codex, a WWW address like the following can be used: http://www.ceec.uni-koeln.de/ceec-cgi/kleioc/0010/exec/pagemed/%22kn28-0083ii_164.jpg%22.

Here the following is applicable:
<collection-reference> = http://www.ceec.uni-koeln.de
<object-reference> = ceec-cgi/kleioc/0010/exec/pagemed/
<granule-reference> = %22kn28-0083ii_164.jpg%22

<**collection-reference**>

In the example, http://www.ceec.uni-koeln.de is obviously a URL. This is where we seek the support of existing library institutions. Obviously a persistent identifier for the individual collections should replace the URL. It would be particularly helpful if the identifiers would incorporate existing schemes for the unambiguous reference to institutions. For example, it would be helpful if the identifier above could be replaced by something that contains a reference to "kn28", the (within Germany) traditional unambiguous reference to the library in question.

Less formal than the rest of these proposals: Within the WWW the question of top-level domains is very much open to discussion, and with "*.museum", at least one type of cultural heritage institution has reached top-level status. Considering the fact that libraries in many ways are *the* nodes of the information network, when we consider the actual amount of information handled, it is fair to wonder if there are there any discussions underway that would lead to the creation of a library top-level domain and references like "www.kn28.de.lib". If not, why not? This could be a very good starting point for persistent implementations and, with the library community directly responsible for the administration of its domains, would do away with an entire level of problems.

<**object-reference**>

Once the problem of the persistency of the basic identifier is resolved, we consider a robust technical solution reasonably simple.

We have implemented the following scheme:

<object-reference> = <interface> <access-mode> <resource-id>

with the following considerations:

<**interface**>

The <interface> of a CEC <object-reference> is a series of one or more identifiers separated by slashes. They represent a software system existing at a given point in time, in our example: kleioc/0010.

Notes:

The reference to a specific interface may be seen as directly opposed to locator persistency. It has been included based on the following assumptions, however:

1) The only thing about which we can be reasonably sure regarding the further development of net-oriented information access is that it will develop considerably beyond the current stage. It is very likely, therefore, that future access systems to digital resources will make them accessible in new ways.
2) On the other hand, it is unlikely that a typical preserving institution will make fundamental changes to its software platform more regularly than, say every ten years.

We would therefore assume that a given institution, when exchanging a software platform 'x' with a software platform 'y' will provide scripts or their future equivalents that will direct all references to the software interface 'x' to methods provided by the new software which closely resemble the previous picture, while at the same time a reference to the new interface 'y' can be provided, making full use of any additional capabilities the new software provides.

As such changes will be infrequent, it is not unreasonable to ask an institution to provide the level of legacy support represented by such scripts.

<**access-mode**>

Almost all imaginable systems for the administration of digital objects will provide access to them according to different qualities, resolutions, access privileges and the like. The <access-mode> of a CEC <object-reference> provides a means to differentiate between different combinations of such properties. Like the interface, it is a series of one or more identifiers separated by slashes.

Notes:

1. It is important to differentiate the type of access granted to an individual object as cleanly as possible from the reference to that object itself.

2. Access-mode notation should, however, not be kept too simple. Relatively soon, standard qualities for digital objects will be developed. In this context, it is important that a mechanism exists that allows the combination of abstractly defined qualities, presumably by standardized names, with specific types of access provided by a library, modeling the peculiar properties of a certain object.

<**resource-id**>

Within the CEC mechanism, a resource-id is a string that allows a direct reference to a specific digital object. A functionally complete set of descriptive data exists so that this object can be accessed (and potentially transmitted) independently of the remainder of the collection. For reasons of better interoperability between resource-ids derived from different collections, we strongly propose that an abbreviated form of collection identifier is contained within the resource-id. This should make it possible to construct a complete reference to a digital object from the resource-id alone.

<**granule-reference**>

Within the CEC mechanism, a granule-reference is a string that allows a direct reference to the smallest division of digitized information within a digital object. Typically this will be the file containing a scanned page. For reasons of better interoperability between references derived from different collections, we strongly propose that the complete resources id is contained within the granule-reference. This should make it possible to construct a complete reference to a digital object from that reference alone.

<**direct-granule-reference**>
A direct granule reference consists of a string that can be used directly to access digitized information on a specific server. It may be necessary to break the reference up into components that represent different levels of a storage hierarchy, and/or into components that map logical names unto physical storage addresses. It does *not* allow for any conceptual interpretation, however. A collection guarantee indicates that the <direct-granule-reference> of a digitized page or other atomic unit of digitization will never change throughout its existence. In our example, kn28-0083ii_164.jpg is a direct-granule reference.

<**mapped-granule-reference**>
A mapped granule reference consists of a string that is separated by a dividing character. The CEEC implementation of the CEC concepts uses a vertical line "|". The first of the two parts is the identifier of a mechanism that allows the second part of the string to be mapped to a direct granule reference, according to a specific set of rules, which may be changed over the lifespan of the digital object or, indeed, be dropped as obsolete. If a mapped granule reference *starts* with the vertical line, it maps to a default mechanism that will exist for the complete lifespan of the object and is called a "canonical reference".

In our example: |kn28-0083ii_82r will map to the file which represents page *82 recto* of the manuscript according to the canonical references given in the literature referring to it. Miller|kn28-0083ii_insertion4-3r may map to the file containing page *3 recto* of the fourth insertion into a hypothesized original manuscript proposed to be assumed by researcher Miller. This interpretation may be adapted according to the researcher's progress or, indeed be dropped if he or she turns out to be mistaken.

Recommendation: *To make digital repositories useful within digital publications, each digital repository should include a publicly accessible interface for access to its component items, which provides a persistent mode of quoting the content of that collection*.

5. Digital collections as integrated reference systems

Thesis: *Traditional libraries support their collections with reference material. Digital collections need to find appropriate models to replicate this functionality*.

Discussion:

The reading room of a manuscript collection traditionally contains sets of reference works that are either very general in nature or relate to the specific collection. Reference materials have to be supplied, in principle, alongside the primary digitized material, as well as to support use of the material. The CEEC site provides an example of a nucleus of such reference literature being integrated into the environment of such material.

Recommendation: *In the next stage of funding, it would be worthwhile to analyze how useful digitization projects are that create repositories of classical reference works digitized according to models that optimize their integration into reference sections of other digital collections*.

6. Library and Teaching

Thesis: *The use of multimedia in teaching is as much of a current buzzword as the creation of digital collections. It is obvious that they should be connected. A clear-cut separation of the two approaches is nevertheless necessary*.

Discussion:

The German "retrodigitization" initiative is one of the most broadly-based funding initiatives for the creation of digital material directly useful for research in the Humanities. It has avoided both an overly centralist approach that would favor a few large institutions, as well as a populist approach of digitizing what creates good press coverage first. It has spread digital collections to a remarkably large number of research libraries, and it has provided digital material that is actually significant for research.

It is disappointing, however, that these materials are not used as much within the academic community as they merit. The general reason for this seems to be that too much of the discussion about digital collections has taken place within the library system. Too few direct links to the appropriate *fora* of the research and educational community have been created. For example, digital library projects seem to be totally absent in the huge recent initiatives in support of multimedia teaching in the university system. More specifically, in the first wave of projects funded by the BMBF (the German federal ministry of education), only *one* Humanities project has been funded, and that project does *not* build upon the digital materials provided by the system of digital libraries. Rumor has it that this will change in the second wave, when at least one such project will presumably be represented, but this seems still to fall much below the level of what should be achieved.

Not to be misunderstood: The author thinks that digital libraries, like conventional ones, should make books and other publications accessible, not write them. And indeed he thinks that a certain tendency to sell selections of highly polished small subsets of material as digital libraries, when they should be classified rather as digital library expositions, has been detrimental to the overall acceptance of digital collections within serious research.

But still, the existing digital collections could be seen as a perfect platform on which to build teaching systems. It would bind multimedia-based teaching to research, as the German tradition of academic teaching has always required. Conscientious efforts towards that goal are needed.

Recommendation: *The possibility of connecting multimedia teaching projects directly to the platforms provided by the existing digital collections should be explored systematically*.